Predicting Product Purchase using Linear Classification Algorithms

K. Maheswari, K. Ponmozhi

Abstract*: *The customer buys the product based on many factors. There is no adequate and properly defined logic for such matter. The customer must satisfy when they see their product itself. They have to trust its quality, price, lifetime of the product, no side effect behavior, name of the product, packing of the product and finally cost. These factors may vary time to time, day to day and even sec to sec. The competition among sellers is also increasing day by day. The choice of choosing the product for customer is more, confused and risky also. Establishing a good relationship among seller and buyer will increase the customer. The retaining of customer is a challenging task. To solve this problem, a model is developed using machine learning algorithms svm, Naïve Bayes, Logistic Regression and fisher's linear discriminant analysis. This model predicts the buying habit of a user/customer. The classification is performed on product purchase dataset and its performance is compared to find which algorithm performs well for this particular dataset. This work is implemented in R software.

***Keywords* : Classification, Logistic Regression, Machine Learning and Naïve Bayes.**

I. INTRODUCTION

Huge amount of data is available in the internet and social media in various forms (unstructured form). There is a need for converting these data in to meaningful, structured and useful one. Data Mining is used to extract information. It mines knowledge from the data. It also performs data cleaning, Data Integration, Data Transformation, data visualization, pattern identification and data evaluation. The knowledge can be mined using some functions. They are Characterization, Discrimination, Association and Correlation Analysis, Classification, Prediction, Clustering, Outlier Analysis and Evolution Analysis. The data analysis may be in the form of classification or prediction. Classification is a technique suitable for categorical values whereas prediction is well suited for continuous values. Lot of classification algorithms is available today. It is difficult to say that, one is superior to the other. The efficiency of the algorithm is based on the dataset taking in to account. The linear classifier algorithms logistic regression, Naïve Bayes, Fisher's Linear Discriminant and SVM are implemented in this work.

Feature selection algorithm Boruta[10] was applied to select important attribute from the dataset. The dataset contains the following attributes of users.

- User ID
- Gender
- age
- Estimated salary and
- Purchased.

The result of Boruta algorithm is shown in figure 1. Boruta performed 12 iterations in 3.929221 secs, 1 attributes confirmed important - purchased, 2 attributes confirmed unimportant - gender, userid; The feature selection result is shown in Fig 1.

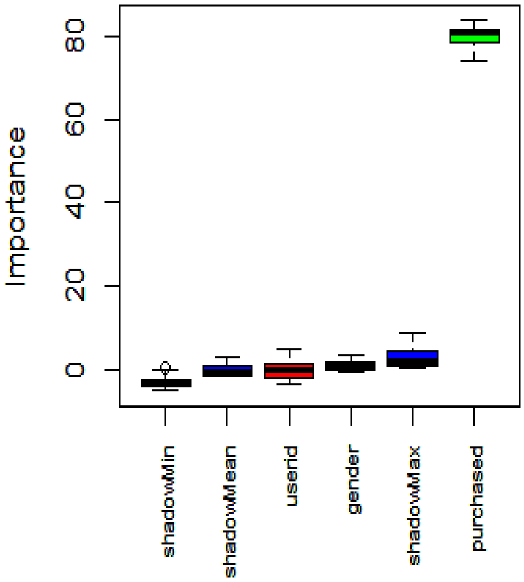

Fig. 1. Feature Selection

The objective is to predict and construct a model for buying the product. The dataset has 400 observations and 5 attribute which describes the details of the website user. The dataset was preprocessed to perform classification models. At last, the performance of the models was measured and compared. The dataset consists of 400 samples which splits into 70% training set and 30% test set. There are 5 attributes and 280 samples in the training set. The test set contains 5 attributes with 120 samples. The Boruta algorithm consists of following steps:

Revised Manuscript Received on December 05, 2019.

* Correspondence Author
Dr.K.Maheswari *, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankovil-Virudhunagar , India. Email: maheswarisnr@gmail.com
Dr.K.Ponmozhi, , Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankovil-Virudhunagar Email: chezhiyan71.p@gmail.com

Retrieval Number: D11121284S219/2019©BEIESP
DOI: 10.35940/ijrte.D1112.1284S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Algorithm 1 : Boruta Feature Selection

```
Input :    Dataset with more samples and attributes
Output : Important features identified for performing
           classification.

1. The features used can be shuffled to remove correlations.
2. Run a classifier algorithm random forest to compute Z
    scores.
3. Compare the Z score of shadow attributes with attributes in
    the dataset. Find the max Z score attribute.
4. Identify Important and unimportant attribute.
5. Consider important attributes for further classification.
6. Shadow attributes generated by this algorithm can be
   removed.
7.The same steps can be repeated for all the attributes is
   assigned as important.
```

The dataset used for this work is extended by adding all attributes. The data is extended by adding at least 5 shadow attributes. The above process will not consider the number of variables in the original set is less than 5. This paper is organized as follows. Chapter 2 describes Background study to develop this paper, chapter 3 focuses the methodology used in this paper, Chapter 4 concentrates results obtained after applying classification algorithms and chapter 5 concludes the work.

II. BACKGROUND STUDY

Data collection related to product purchase is very important task. Most of the large scale business people are moving towards machine learning for their promotion. Predicting the demand of a product is a challenging task in on line business. To meet the challenge, there is a need for reviewing historical data, methods, procedures, techniques, suggestions and findings of various authors currently working on that area.

The author [1] reviews e-commerce product using SVM. A new algorithm IG-RS-SVM was proposed for text classification. Information Gain algorithm was used for feature selection. The experimental results show that the IG-RS-SVM performs effectively and improved accuracy. Ling Tang et al.[2] described a hybrid model for online purchasing to predict whether the customer is going to make a purchase or not. Firefly Algorithm (FA) was proposed to predict accuracy. Amy Li et al.[3] analyzed amazon.com's product for classification. SVM and kernel methods were implemented to predict the product to be co purchased or viewed. The author considered textual title and category for classification. The high degree of accuracy is achieved as results.

Kanako Komiya et al.,[4] proposed a new model negation naïve Bayes to categorize products on a web. The author compared NNB with the NB and CNB. The NNB outperforms significantly. Gajanan P et al., [5] considered the parameters users session, history generation and user activity. Naïve Bayes algorithm is applied on clickstream data. The analysis was made in customer relationship management and business intelligence. Barbu Bogdan Popescu [6] demonstrated multivariate discriminant analysis in bank marketing. This work concentrates the increase of sales by attracting more customers.

Pujari C et al.,[7] a new frame work to identify customer opinion was designed. Three classification algorithms were used for classification (Naïve Bayes, Maximum Entropy Classifier and SVM). The performance of these three algorithms was analyzed. Abhinav singh et al.,[8] suggested customer review platforms and decision support systems as a data source. This helps to increase the chance of product sale in a successful way. The accuracy of 82.44% was achieved by classification algorithms. Upma Kumari{9] et al., created algorithms to classify smart phone product review. The author used various datasets to classify sentiment sand texts. Three performance features precision,recall and f-measure was considered and accuracy is calculated. The author concluded that the obtained and predicted results shown high accuracy.

Miron B et al., [10] developed a new model for feature selection. The author demonstrated the important and unimportant behavior using boruta feature selection algorithm. Rohit Joshi et al., [11] attempted with different factors played in online purchase of customers. The author tried for different products for Indian customers. The questionnaire is prepared for survey. The author suggested to retailers to predict purchase behavior based on location.

Discriminant analysis [13] can be done by Fisher's rule for multiple groups to perform classification. The error rate is used for measuring the performance. The expression is derived by the author for the optimal error rate.

This study focused product purchase behavior using different algorithms, different parameters, different dataset with different results. In this work, the models used svm, naïve bayes, logistic regression and fisher's discriminant analysis with age, salary as independent variable and purchased is a dependent variable. The experimental results with comparison of models are also presented.

III. METHODOLOGY

A. Problem Statement

Predicting the customer to buy the product or not, based on their age and salary. Various linear classification algorithms are applied in the dataset and finally accuracy is measured.

B. Framework of the Problem

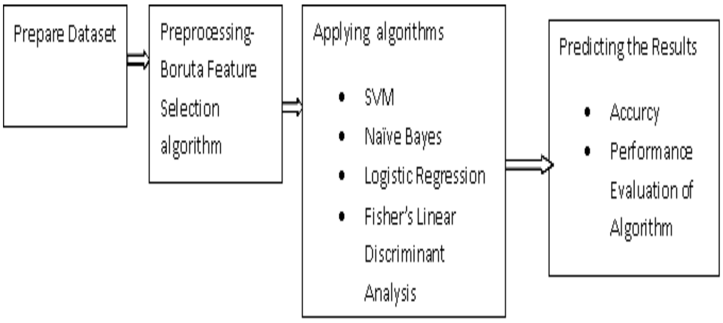
Prepare Dataset
Preprocessing-Boruta Feature Selection algorithm
Applying algorithms
SVM
Naïve Bayes
Logistic Regression
Fisher's Linear Discriminant Analysis
Predicting the Results
Accurcy
Performance Evaluation of Algorithm

Fig. 2. Workflow of the Problem

The social_ network_ads.csv file is downloaded from the github (https://github.com/atse0612/Machine-Learning-A=Z/blob/master/Social_Network_Ads.csv).

The frame work of the problem is shown in Fig 2. The preprocessing was performed to clean the data set, to select important features, minimize number of attributes and reduce the complexity of process. The Linear classification algorithm is used for classification and finally accuracy is measured.

The svm classifier [14] [16] [17] is used for finding the purchase rate of the people. SVM constructs hyper plane for classifying the products. It performs classification for both categorical value and continuous value. For categorical value, it creates some dummy variables such as 0 and 1. Naïve Bayes [15] algorithm is based on probability family. It works based on Bayes' theorem and is stated mathematically as follows.

$$P(m/N) = \frac{P(N/m)\,P(n)}{P(N)} \qquad (1)$$

Where

P(m/N)	posterior probability
P(n)	class prior probability
P(N)	predictor prior probability
P(N/m)	Likelyhood

Features used in the naïve bayes are not dependent to each other. From studies, it is found that, naïve bayes outperforms than other classification algorithms.

Logistic regression is very similar to linear regression and the output value being modeled is 0 and 1. The logistic regression equation is

$$Y = e^{\frac{(a0+a1\,x)}{(1+ e^{a0+a1\,x})}} \qquad (2)$$

Y is the predicted result, a0 is bias, a1 is coefficient for single input value x. It is a binary two class classification algorithm. It is also based on the family of probability. The probability score is always between 0 and 1.

Algorithm 2 : Logistic Regression

Input : Dataset with more samples and attributes
Output : Important features identified for performing classification.

1. Calculate the boundary function using log odds function.

   $$BODD = B_0 + a\,B_1 + b\,B_2 \qquad (3)$$

   BODD Boundary Function or Log odd Function. Linear equation shows the relationship between variable and outcome.
2. Calculate the ratio of odds.

   $$R = e^{BODD} \qquad (4)$$

   R is the ratio of odds function.
3. Calculate logistic function using the following formula.

   $$P = \frac{e^{BODD}}{1 + e^{BODD}} \qquad (5)$$

   P is Logistic function.
4. Repeat the steps 1,2 and 3 until step 3 is achieved.

Fisher's linear discriminant classifies multi-dimensional data. The coefficient of linear equation is used to perform classification for the classes.

Algorithm 3 : Fisher's Linear Discriminant Analysis

Input : Dataset with more samples and attributes
Output : Good class separability

1. Calculate separability between data points. The separability is calculated by finding the distance between different classes.
2. Calculate inner class variance.
3. Construct fisher's criterion to maximize between class variance found in step1 and minimize inner class variance found in step2.
4. Repeat the steps 1,2 and 3 until step 3 is achieved.

IV. EXPERIMENTAL RESULTSTH

The parameters considered in this work are svm-type(C-classification), SVM-Kernel(linear), cost(1) and gamma(0.5) with Number of Support Vectors 116. The confusion matrix for test set is shown in TABLE I.

TABLE I. CONFUSION MATRIX

Prediction	y	pred
	0	1
0	57	7
1	13	23

The Accuracy achieved is 80% which is shown in TABLE II. The kappa statistics compares the relative accuracy of the random system. Its value lies between o and 1. Its value never exceeds 1. The sensitivity specifies that 81% of people will have positive opinion. The predicted results says that, there is a chance of 81% of people will buy a product. The specificity results interpreted as, among 81%, 75% will have negative opinion.

TABLE II. SVM ACCURACY

Parameters	Prediction
Kappa	0.5495
Mcnemar's Test P-Value	0.26355
Sensitivity	0.8143
Specificity	0.7667
Pos Pred Value	0.8906
Neg Pred Value	0.6389
Detection Rate	0.5700
Detection Prevalence	0.6400
Balanced Accuracy	0.7905

SVM classifier is plotted and is shown in Fig 3. The x in this model represents support vectors and o represents other data points.

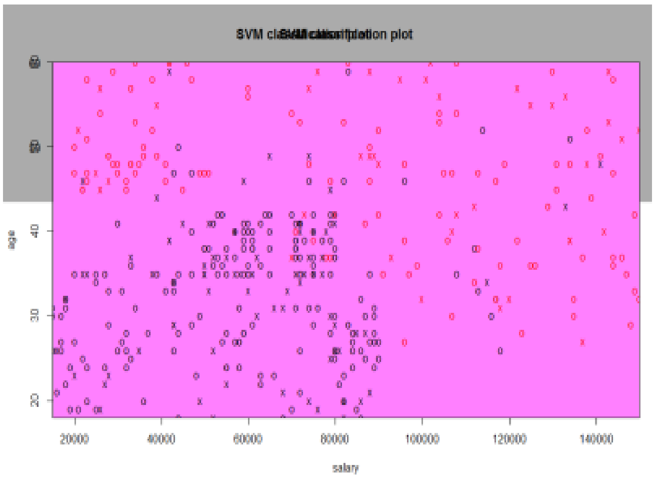

Fig. 3. SVM Classification

The above result is obtained for linear svm. The person whose salary is more than one lakh and age is above 40, that kind of people will make a purchase. In order to achieve higher accuracy further, the svmt model is tried with different attributes. The number of support vector is 153, with gamma=0.2 and cost = 100. The dependent variable is purchased and the independent variable is salary and age. The prediction result is shown in TABLE III.

TABLE III. SVM PREDICTION

Prediction	y	pred
	0	1
0	0	1
1	100	0

The model is plotted and is shown in Fig 4. The age and salary is normalized because age is in the range of 20 to 80 but salary is 5 digit number starts from 20000 to 140000. So, bringing both the values in the same format is required. After scaling, the performance was not improved.

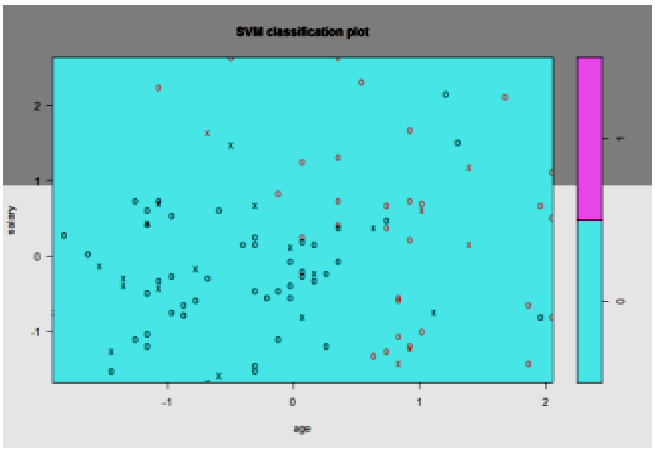

Fig. 4. SVM Classification with different attribute

In the above plot,. The positive values in the graph shows positive predictions and negative value shows negative prediction. The confusion matrix is in TABLE IV.

TABLE IV. CONFUSIONMATRIX FOR DIFFERENT ATTRIBUTE

y_pred	0	1
0	64	0
1	36	0

The Accuracy achieved is 0.64, first SVM model shows 80% accuracy and the second model shows only 64% accuracy. so, the next algorithm naïve bayes is implemented.

The data modeling begins with data splicing. The dataset is split with train set and test set. Dividing train set and test set of the dataset is known as cross validation. The trainset is the part of the dataset to build and train the machine learning model. The test set is used to evaluate the prediction that was performed. The ratio of training set and test set is 70:30. The accuracy of train prediction is 85% and test set is 65% which is shown in fig 5 and fig 6.

The naïve bayes calculates prior probabilities. The True Positive (TP) rate is the amount of instances that are classified as class y, The confusion matrix given above says that, out of 257 instances 239 is correctly predicted, and 18 are incorrectly predicted.

From the confusion matrix, 99 people will have an idea of buying product. the correctly predicted result from the dataset to buy a product is 99 and not buying a product is 239.The accuracy is 85%. The train prediction value and test prediction value is shown in TABLE V and TABLE VI. The train accuracy and test accuracy is shown in Fig 5 and Fig 6.

TABLE V. TRAIN PREDICTION RESULT

Train	y	pred
prediction	0	1
0	239	18
1	44	99

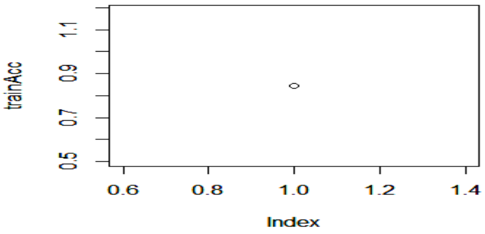

Fig. 5. Train Accuracy

TABLE VI. TEST PREDICTION RESULT

Test prediction	y	pred
	0	1
0	64	0
1	36	0

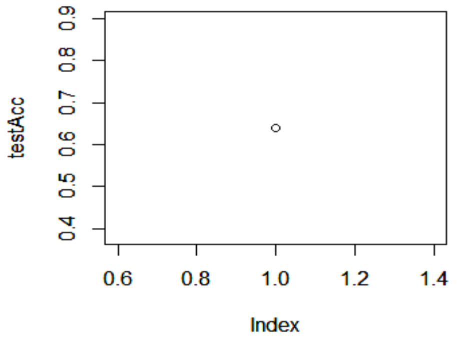

Fig. 6. Test Accuracy

Test accuracy result is 65%. There is no rule or law to say that the test set accuracy *has to be* lower than the training set accuracy. Once the accuracy is not improving, we can stop training. The next model logistic regression is going to be implemented. Logistic regression is an extended version of linear regression. It is mostly used for predictive analysis. It shows the relationship between one dependent attribute and one or more other independent attribute. The prediction value is shown in TABLE VII.

TABLE VII. PREDICTION RESULT OF LOGISTIC REGRESSION

prediction	FALSE	TRUE
	0	1
0	239	18
1	46	97

Its prediction may be binomial which takes the value 0 and 1. The accuracy is calculated by the following formula

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives). (97 +239) / (239 + 18+46+97) (336 / 400) = 84%. The true positive and true negative values are shown in Fig 7.

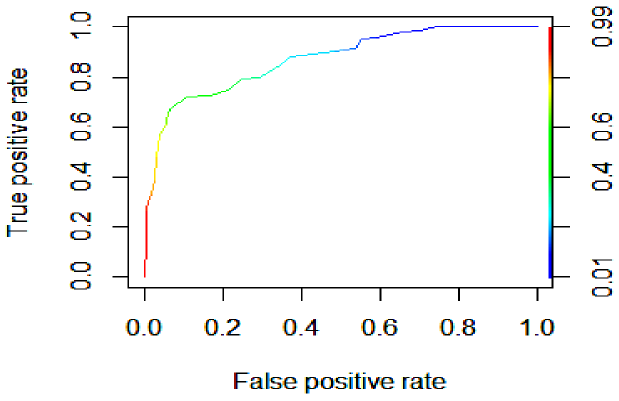

Fig. 7. Accuracy

Only 97 people will have an idea to buy a product. The line in the graph shows the outcome measures are varying for varying threshold values. Each point in the plot represents specificity and sensitivity. The curve passes through in the upper direction only. The highest accuracy towards specificity and sensitivity is achieved. The error rate is 16%.

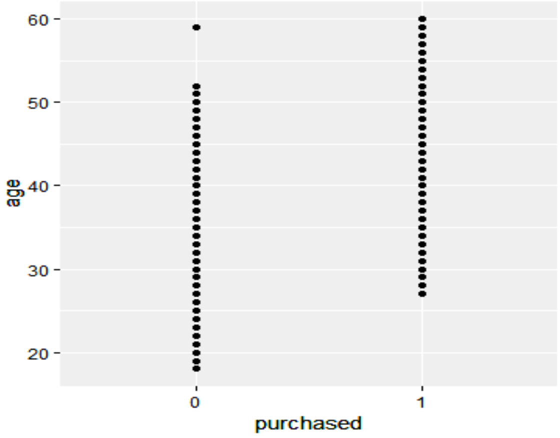

Fig. 8. worthless Test Accuracy

In Fig 8, there is a overlap of age from 25 to 50. Some people of age from 25 to 50 will not buy a product and some people of age 25 to 30 will buy a product. Sometimes this is called worthless test. We cannot predict correctly whether they will buy or they will not buy. It indicates 50% specificity and 50% sensitivity.

LDA method use mean and variance for classification. It creates linear boundary between mean and variance. There are two different models that are developed in this work. First model depends on the variable purchased and the second model depends on age and salary. The response class is classified by the LDA classifier algorithm. The x axis represents the coefficient value of LDA model. The Prior probabilities of groups are presented in TABLE VIII.

TABLE VIII. PRIOR PROBABILITIES OF GROUPS

Group 0	Group 1
0.6425	0.3575

The group means and coefficient of LDA is shown in TABLE IX and TABLE X.

TABLE IX. GROUP MEANS OF LDA

Group Means	AGE	SALARY
	0	1
0	32.79377	60544.75
1	46.39161	86272.73

TABLE X. COEFFICIENTS OF LDA

Coefficients of LDA	AGE	SALARY
	1.109230e-01	1.598719e-05

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The good practice is always comparing the results of various analysis methods. This comparison helps to verify the results obtained and bringing new ideas, assumptions, opinions, behaviors and characteristics.

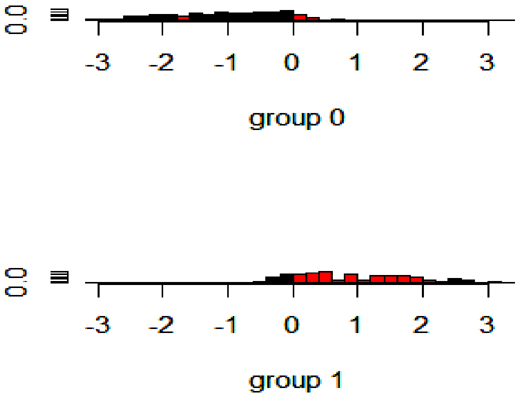

Fig. 9. First Discriminant Function

The separation part and overlapping part of groups after predicting classes is shown in Fig 9. There is a small overlap between group 0 and group 1. This function calculates the ratio of separation for within group and between group variances. The First Discriminant Function itself separates group 0 and group 1 as well. To increase the separation again, the second discriminant function is performed which is shown in Fig 10. The same level of separation is maintained in both the function.

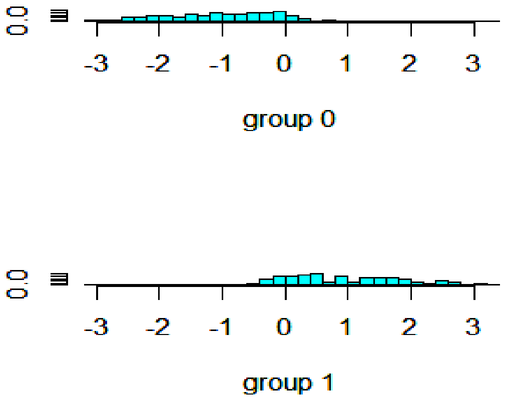

Fig. 10. Second Discriminant Function

The prior probabilities of training data 64.25% evaluated as 0 and 35.75% is evaluated as 1. 0 represents non buying behavior and 1 represents buying behavior. From the group means, the people those who have age above 35 and salary nearly 80000 is predicted for buying products.

Our model contains two different classes age and salary. So, the boundary between the classes is calculated by the following formula 1.109230e-01 * age +1.598719e-05 * salary. The coefficient of first variable age is greater than the second variable salary. It is concluded that the first variable influences more to purchase the product than the second variable.

The partimat() is a function to plot linear discriminant. This function is used under the installation of klaR package. The linear and quadratic classification of two features can be displayed at a time. The linear or quadratic classifications is plotted in Fig 11.

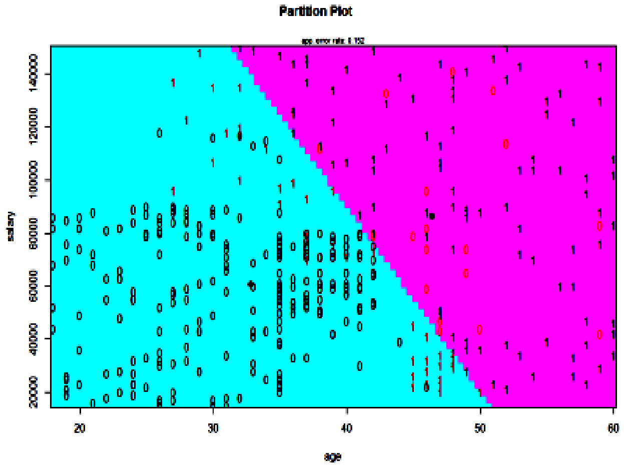

Fig. 11. Linear or Quadratic Classifications

The algorithms svm, Naïve Bayes, Logistic Regression and fisher's linear discriminant analysis were applied and compared its performance which is shown in Fig 12. Naïve Bayes shown highest accuracy for this dataset. Almost all the algorithms shown the accuracy above 80% except fisher's Linear Discriminant Analysis.

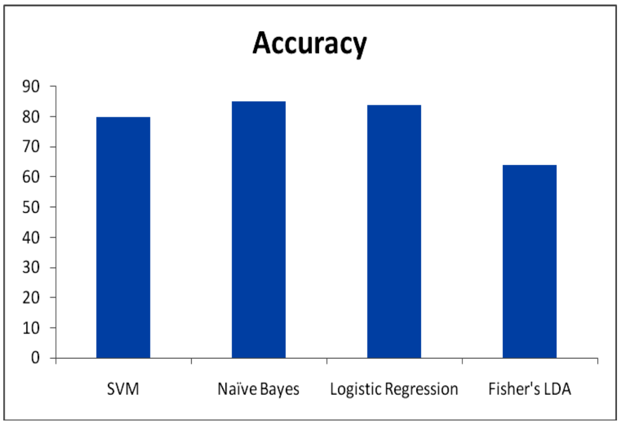

Fig. 12. Performance of Algorithm

V. Conclusion

Never-ending factors are available to examine the customers buying behavior. Seasons, weather, dynamic price will also influence the sales. A well designed dataset and algorithm is needed to solve this problem. The choosing of dataset takes more time. 80% of time will spent for preparing or finding the dataset. The incomplete and inaccurate data will give negative opinion. Most of the survey says that, 26% of the dataset is not proper. The main contribution of the work is improving the accuracy in terms of age, and salary. The future work will move towards other advanced algorithms.

REFERENCES

1. Jiajun Ye a , Huan Ren b and Hangxia Zhou c “An IG-RS-SVM classifier for analyzing reviews of E-commerce product”, International Conference on Information Technology and Management Innovation (ICITMI 2015).
2. Ling Tang , Anying Wang , Zhenjing Xu , Jian Li,” Online-Purchasing Behavior Forecasting with a Firefly Algorithm based SVM Model Considering Shopping Cart Use”, EURASIA Journal of Mathematics, Science and Technology Education ISSN: 1305-8223 (online) 1305-8215 (print) 2017 13(12):7967-7983 DOI: 10.12973/ejmste/77906
3. Amy Li, Mitchell Hughes, Nolan Hodge,” Classification of Amazon Product Metadata: Support Vector Machine & Kernel Methods”, March 26, 2018.
4. Kanako Komiya , Naoto Sato , Koji Fujimoto , Yoshiyuki Kotani ,” Negation Naive Bayes for Categorization of Product Pages on the Web”, Proceedings of Recent Advances in Natural Language Processing, pages 586–591, Hissar, Bulgaria, 12-14 September 2011.
5. Prof. Gajanan P. Arsalwad , Angela A. Dhanawade , Rupali J. More , Pranali A. Kulkarni,” Analysis of E-Customers Behavior Using Naïve Bayes Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2017.
6. Barbu Bogdan Popescu,”using the Method of Discriminant Analysis in Bank Marketing”, International Journal of Business, Humanities and Technology, Vol. 5, No. 3; June 2015.
7. Pujari C., Aiswarya, Shetty N.P. (2018) Comparison of Classification Techniques for Feature Oriented Sentiment Analysis of Product Review Data. In: Satapathy S., Bhateja V., Raju K., Janakiramaiah B. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 542. Springer, Singapore.
8. Abhinav singh, Conrad S.Tucker,” A machine learning approach to product review disambiguation based on function, form and behavior classification”, Decision Support System , Elsevier,ScienceDirect,Vol 97,May 2017,PP:81-91.
9. Upma Kumari,Arvind K.Sharma,Dinesh Soni,” Sentiment analysis of smart phone product review using SVM classification technique”,International Conference on energy,communication,Data Analytics and Soft Computing,IEEE,2017.
10. Miron B. Kursa, Witold R. Rudnicki,” Feature Selection with the Boruta Package”, Journal of Statistical Software, September 2010, Volume 36, Issue 11.
11. Rohit Joshi, Rohan Gupte, Palanisamy Saravanan, A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers, published by Theoretical Economics Letters,Vol.8 No.3, February 2018.
12. Filzmoser P., Joossens K., Croux C. (2006) Multiple group linear discriminant analysis: robustness and error rate. In: Rizzi A., Vichi M. (eds) Compstat 2006 - Proceedings in Computational Statistics. Physica-Verlag HD
13. Daniel J. Tozera,4, Gerard R. Daviesa , Daniel R. Altmanna,b , David H. Millera , Paul S. Toftsa,” Principal component and linear discriminant analysis of T1 histograms of white and grey matter in multiple sclerosis”, Magnetic Resonance Imaging 24 (2006) 793 – 800,Science Direct.
14. Dr.K. Maheswari, P. Packia Amutha Priya , "Analysis and Implementation of Text Mining for Different Documents", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 3, Issue 5, pp.109-113, May-June-2017.
15. Dr.K.Maheswari,Ms.P.Packia Amutha Priya “Classification of Twitter Data Set using SVM and KSVM”, Published in international Journal of Pure and Applied Mathematics, Volume 118 No. 7 2018,PP: 675-680, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) , Scopus Indexed.
16. K.Maheswari,“Improving Accuracy of Sentiment Classification Analysis in twitter Data Set Using knn” published in International Journal of Research and Analytical Reviews,Vol 5, Issue 1, PP:422-425/E ISSN 2348-1269 Print ISSN 2349-5138,UGC Approved Journal.
17. Dr.K.Maheswari,Ms.P.Packia Amutha Priya “Predicting Customer Behavior in Online Shopping Using SVM Classifier”, presented paper in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization & Signal Processing, INCOS'17, published in IEEE Xplore 01 March 2018.

AUTHORS PROFILE

Dr.K.Maheswari received her B.sc (Computer Science) from Madurai Kamaraj University and MCA. M.Phil. from Bharathidasan University. She has completed her Ph.d at Bharathiar University. She is currently working as an Associate Professor in the Department of Computer Applications, Kalasalingam Academy of Research and Education. She has 23 years of teaching experience. She has presented research papers in several national and international conferences. She has published many research papers in various international journals. Her research interest is VoIP , network security and Data Mining.

Dr.K.Ponmozhi is currently working as Assistant professor in the Department of computer applications, Kalasalingam Academy of Research and Education, Srivilliputhur, TamilNadu, India. She received her Bachelors degree in computer science from Madurai Kamaraj University, her MCA from Bharathidasan University. She received her Ph.D in computer science from Mother Teresa women's University. She has more than 20 years of teaching experiences. She published many articles in journals and conferences both in National level and International Level. Her area of interest is networking, computer graphics, image processing, cognitive science.